CHAPTER TWO

TESTS: USES, TYPES AND VALIDATION

BY
OKE, T.D.

DEFINITION OF TEST

The Penguin Dictionary of Psychology defines test as:

a standardized type of examination given to a group or individuals; it may be qualitative or quantitative i.e. determine the presence or absence of a particular capacity, knowledge or skill, or determine the degree in which such is present: in the later case, the degree may be determined by the relative position of an individual in the group or the population, or by assigning a definite numerical value in terms of some selected unit.

This definition suggests that the purpose of a test is to identify or discover what a person can do under certain controlled circumstances.

Test according to Adekunle (2003) is an instrument used for providing a description of a person and his attributes so that:

i. Prediction could be made;
ii. Inter and intra unit comparison could be made.

Anastasi (1968) defines a test as an objective and standardized sample of behaviour. A test is said to be a sample of behaviour because, from the test performance, an inference is made regarding the testee's over all behaviour in relation to the issue tested.

Makinde (1983) assumes that a test is generally a set of questions, problems, puzzles, symbols, and exercises used to determine a persons ability, aptitude, knowledge, qualifications, interest and level of social adjustment. To Nwoye (1990), a test is a set of stimuli presented to an individual in order to elicit responses, on the basis of which a numerical score can be assigned.

According to Ibanga (1981), a test is a systematic procedure for measuring a sample of behaviour. By systematic procedure, it means that a test is constructed, administered and scored according to prescribed rules.

The behaviour in the definition suggests that a test measure only the test taking behaviour, that is, the responses a person makes to the test

items. This means that a person can not be measured directly, rather his characteristics from his responses to the test items are inferred.

A test contains only a sample of all possible items because it is not possible to set a test to cover every item that might be developed to measure human traits.

USES OF TESTS

The functions of educational tests are numerous and varied. Obe (1980) classified these functions as follows:

i. Instructional and motivational aids
ii. Evaluation of scholastic achievement for promotion, certification etc.
iii. Selection and counselling
iv. Research for evaluating curricular, teaching methodology and effectiveness.

Ibanga (1981) cited the following as the most frequent uses of tests:

1. To express quantitatively several significant characteristic of pupils abilities, achievements, aptitudes and interests.
2. To group students for the purpose of instruction.
3. To plan special instruction for children with identifiable strengths or weaknesses.
4. To identify students with certain characteristic – e.g. the under-achiever, the poor reader, the fast learner, etc.
5. To evaluate progress of individuals or groups and compare classroom achievements.
6. To adapt instruction to the ability level or educational background of the individuals or groups.
7. To secure information on the range of talent within a group.
8. To identify children whose potential or whose achievement seems to be irregular from year to year.
9. To plan instruction to build upon the present level of achievement-avoiding duplication of what has been mastered and calling attention to areas where review or remedial work is indicated.
10. To identify children whose education and vocational goals are not in harmony with their ability.
11. To identify children to be referred to the school counsellor.

12. To help children to understand their own abilities.
13. To provide information to student as to their relative chances of success in various courses or vocations.
14. To help parents to understand their children.
15. To compare growth from year to year.
16. To compare levels of achievement among several schools in a school system or between a school system on a national norm.
17. To identify curriculum areas that may need study or revision.
18. To gain insight into the effectiveness of instruction and curriculum planning.
19. To improve supervisory practices.
20. To stimulate teachers to greater awareness and understanding of individual differences among their pupils and to increase their capacity to apply this awareness and understanding to an improved school programme.

Anikweze (2005) pointed out the following as some of the uses of tests to the classroom teachers.

i. Tests are used in making objective observations of the learners. That is, when correctly applied, tests lend precision and objectivity to teachers impact on the learners changes in behaviour.
ii. Tests are used to measure educational gains consequent upon some instruction.
iii. Tests are used to assist teachers pace their instruction having assessed the extent to which learner have mastered any piece of content before the teacher proceeds to another content area.
iv. Good testing instruments are useful in discovering both the unseen and un-seeable potentials in learners as well as their deficiencies.
v. Tests are used to elicit behaviour from learners and sample their performances.
vi. Tests motivate learners to study better because tests act as stimulus for learning and most students are gingered into studying whenever tests are proposed.
vii. Tests are useful for administrative decisions in education e.g. ability streaming, selection and certification.

In summary, the purposes of educational tests are:

1. To determine the extent to which students have benefitted from a

course of instruction.

2. To satisfy ourselves as teachers that the methods of teaching adopted are effective.
3. To diagnose students' strengths and weaknesses.
4. To help in predicting students' future performance.

COMMON DIMENSIONS FOR CLASSIFYING TESTS

There are so many ways of classifying tests depending on the functions/purposes. structures and methods.

1. CLASSIFICATION BY METHODS

a. Individual versus group tests: A test which is designed to test only one candidate at a time is called individual test and may be either performance or oral test as in the case of interview. While those which can be administered to more than one person at a time are called group tests. Group tests are usually pencil and paper tests. Example is the end of semester examination.

b. Speed versus power tests: The test items are usually very simple in speed test, but there is a stringent time limit. They are designed to find out how many items of the test. a student can cover within specified time limits; while power tests are composed of items of varying difficulty and have a time limits that allows completion of all items. They are designed to find out how far a student can demonstrate the depth of his knowledge and understanding on some given number of problems. Thus, the score reflects the level of difficulty of item that the test taker can answer.

c. Alternative versus free response tests: Alternative response item requires the test taker to select the most appropriate response from among several alternatives, as in multiple choice, true-false, or matching items. On the other hand in free response item, the testee is required to supply a response as his own personal answers to the questions posed as in completion, short answer or essay.

d. Maximal versus typical performance tests: Maximal performance tests require the testee to attain the best score he possibly can, generally, achievement and ability tests are examples of maximal performance measures. Whereas, typical performance tests seek to measure the students usual or habitual performance.

Personality tests are examples of typical performance measures.

e. **Paper-and-pencil versus performance tests:** Paper-and-pencil tests are those designed to test for growth in the different levels of cognitive functioning. Usually test items are presented and responses made on paper. Performance tests are those normally designed to test for competence and expertise in technical and psychomotor areas, like in Fine Arts, Theatre Arts, Physical Education etc which involve manipulation of some objects or apparatus.

f. **Culture free versus culture fair test:** A culture fair test in both its content and language, no testees from a particular cultural group are given undue advantage. A good example of culture fair test is the intelligence tests of R.B. Cattel which contain short non-verbal tests that use pictures and diagrams that are of common meaning to a wide variety of cultures.

A culture-biased test is a test whose main items contain concepts and materials that are typically familiar to people of a select culture for whom the test is intended.

An opposite of culture-biased test is culture free test which is the test whose test item contents do not reflect any particular cultural setting (Nwana, 1981) culture free tests, are free from the cultural characteristics of any particular human community. The Raven's Progressive Matrices is a good example of culture free tests.

2. CLASSIFICATION BY PURPOSE

a. **Achievement test:** This measures learning that has occurred. Such learning may result from experiences in a relatively circumscribed learning situation as a classroom or training programme. The frame of reference is on what has been learned. Most of the tests that students have been taking in their various institutions are achievement tests.

b. **Aptitude test:** It measures the ability of a student to acquire or learn certain behaviours or skills if given appropriate training. Its frame of reference is toward the future. Aptitude test is designed to measure capacity to learn, that is, to predict what one can accomplish with training.

c. **Ability test:** This measures the power to perform a task. It is used to evaluate individual differences in skills and knowledge, achievements and aptitudes. Ability tests measures the results of

more general or broad learning experiences.

3. **CLASSIFICATION BY STRUCTURE**

a. **Essay test:** The essay type is the traditional test in which the examiner supplies the questions while each examinee composes his answers in the form he pleases. Choice questions are often provided so that all the testees do not necessarily have to answer the same set of questions.

b. **Objective test:** This test requires an examinee to answer many questions in a short time in a structured way either by underlining, encircling or shading the letter corresponding to the correct answers. There are no choice questions, and each question usually carries only one mark, there are many different types of objective tests, the commonest four being the multiple-choice, true-false, matching and the completion (or short answer). These tests are called objective because similar answers by different testees are given the same marks, no matter who did the scoring, provided the scoring key is correct.

c. **Oral test:** The oral type is particularly useful as an instructional aid. Oral questioning during classes ensures immediate feedback and students' participation in class discussions, thus aiding their mental alertness. The oral test is employed at interviews, language examinations, and as viva voce in professional examinations such as Ph.D defence of thesis. In the Nursery, Kindergarten and lower Primary education, the oral examination is frequently used to evaluate achievement because of the pupils' inability to read and write well.

STANDARDIZED AND TEACHER-MADE TEST

There are many types of tests but two types are frequently used for measuring educational performance. These are the standardized tests (very rarely used at present in Nigeria) and teacher-made tests (very widely used in our school system).

STANDARDIZED TESTS

It is a test in which the procedure, apparatus and scoring have been fixed so that precisely the same testing procedures can be followed at different times in different places (Anikweze, 2005).

Obe (1980) defines standardized test as one for which there are:

i. The same set of questions for all candidates.
ii. Specific procedures for test administration and scoring.
iii. Specific method of score-interpretation, and
iv. Test data about norms, reliability and validity.

Most psychological tests are standardized e.g. The Stanford-Binet intelligence scale, the Wechsler intelligence scale etc standardized tests are generally said to be norm-referenced because they are designed to show how an individual performs as compared with others taking the same test.

MERITS OF STANDARDIZED TESTS

i. The tests are objective.
ii. All testees are evaluated by the same yardstick under similar conditions.
iii. If used throughout a student's school career they can provide systematic longitudinal information.
iv. They are very useful for national or statewide comparison of students.
v. Provide norms for a wide geographical area provided all the schools involved adopt the same syllabus.

DEMERITS OF STANDARDIZED TESTS

i. They may not cover certain objectives of the local school.
ii. The tests may control what is taught instead of the latter controlling the former.

TEACHER MADE TESTS

It is a test constructed, administered, and scored by the classroom teacher, or possibly by a committee of several teachers in the same school (The mid-semester, final semester, terminal and promotion examinations are usually teacher made tests.

Two major categories of tests are used by teachers for assessing the achievement of their instructions and for differentiating and certifying their pupils. These are free response otherwise known as essay tests and structured response or objective tests.

ESSAY TESTS

An essay test measures the non-structured types of learning such as creative writing, critical thinking, problem solving, imagination and organizational ability. Essay test could be free response or restricted response. Many essay type tests according to Anikweze (2005) have typical features that distinguish them from the structured response test. These include:

a. Each testee answers a small number of questions.
b. Scripts are written in examinees' individual styles.
c. Examinees have the freedom of organizing their answers in their own unique ways.
d. Each testee operates under minimum constraints in presenting the answers.
e. Quite often choice options are available as a way of covering the scope of the instructional content.

Obe (1980) summarises the merits and demerits of essay test as presented below:

S/N	MERITS	DEMERITS
1	Easy to test.	Making is tedious and time consuming.
2	Measures writing, expression, memory organization, problem solving and originality.	Marking is often subjective, susceptible to 'halo' effect and unreliable.
3	Measures depth of knowledge in a restricted area.	Narrow coverage of Instructional Objectives.
4	Useful in many subject areas.	Choice of questions renders the test poorly standardized
5	Can serve as projective test.	Unfair to slow writer and those not verbally inclined.
6	Minimises guess work and cheating.	Test administration is often long, Boring and anxiety- provoking.
7	Make students prone to cram-pour-forget syndrome.	
8		Prone to leakage.
9		Encourages bluffing.
10.		Consumes answer papers.

OBJECTIVE TESTS

The objective test is the most popular type of test used in Nigeria. Objective test items include:

a. Completion items which have blank spaces that are to be filled with one or two words or a phrase per item e.g. The capital of Ogun State is _____

b. True-false items, some times called 'Yes' or 'No' answers, require an alternative response e.g. Chief M.K.O. Abiola won the presidential general election of 1993 but was not declared winner (Yes or No) True or False.

c. Matching items require the matching of two or more associated words or phrases according to given directions e.g. nations and capitals, nations and major exports, terms and definitions, dates and events, events and places, events and results, books and authors, causes and effects, inventors and inventions etc. Example: Match each capital in column 1 with the right country in column II.

S/N	COLUMN I	COLUMN II
1	Abuja	a. Egypt
2	Accra	b. Sierra Leone
3	Freetown	c. Nigeria
4	Cairo	d. South Africa
5	Pretoria	e. Liberia
	f. Ghana	
	g. Togo	

d. Multiple-choice items require the students to select the proper answer from a number of possible alternatives e.g. Federal College of Education, Osiele, Abeokuta presently has _____ number of schools (a) 3 (b) 4 (c) 5 (d) 6

CHARACTERISTICS OF OBJECTIVE TESTS

The following are the distinctive characteristics of objective tests:

1. Each candidate chooses and answers from a limited list of options.
2. Each question consists of a stem and options.
3. Each item has a predetermined key which is the correct answer.

4. It permits reliable measurement of an extensive sample of fatual material.

The strengths and weaknesses of objective tests as summarized by Obe (1980) are presented below:

S/N	MERITS	DEMERITS
1	Easy to mark.	Difficult to set.
2	Marking is objective and Reliable.	Poor in measuring expression, organisation, originality and higher thought processes.
3	Wide coverage of Instructional Objectives.	Does not measure deep knowledge.
4	Absence of choice questions makes test better Standardized.	Prone to cheating and giraffing during administration.
5	Fair to all candidates, Slow writers..	Consume question including Papers.
6	Test administration is short and not anxiety provoking	Encourages guessing.
7	Not prone to leakage.	
8	Economical with answer papers.	
9	Promotes fast thinking and adaptable to many subject areas.	
10		Easily refined through item Analysis.

VALIDATION OF EVALUATION INSTRUMENT

The results of educational measurements can be used for many purposes.

Therefore, teachers should be well informed on how to validate testing instruments. Good instrument for evaluation should provide for effective testing. An effective test according to Anikweze (2005) is a test that satisfies certain requirements or properties of a useful instrument as judged by experts in measurement and evaluation. The measuring instruments should possess at least three essential properties; validity, reliability and usability.

TEST VALIDITY AND TYPES

The validity of a test refers to whether the test measures what is intended to measure. It means the truthfulness of a test. It is the extent to which the test serves the purpose for which it has been designed. A test must therefore satisfy the edumetric as well as psychometric functions of evaluation. That is, a test should be able to measure the achievements of the learner, discriminate them according to their demonstrated abilities and at the same time, be appropriate for predicting subsequent outcomes.

Since there are various purposes for which tests can be employed, there are also several distinctive kinds of validity. The most important three being: content validity, criterion-related validity and construct validity.

CONTENT VALIDITY

Content validity is also known as curricular validity, because it is determined by curricular method. It is the extent to which the test measures the subject matter content and the instructional objectives stipulated for a given course. Since, it is not possible for an achievement test to cover everything taught in a course, the emphasis is on test coverage of a "large and representative sample" of the course content.

When an achievement test appears to be good on a cursory glance, it is said to have a good face validity. Though, face validity is necessary but not enough because it is insufficient to certify the condition for content validity.

Content validity demands that all aspects of the syllabus should be adequately represented in the testing instrument. For a test to be said to be valid content wise, it must be guided by a table of specification.

CRITERION – RELATED VALIDITY: This is the extent to which scores obtained with a test are related to some other present or further measures. Two types of criterion – related validity is predictive and concurrent validity.

Predictive validity: This is the test's ability to predict or estimate or forecast students future performance on a given task. For instance, continuos assessment scores in EDU 223 may be used to predict students performance in examination. The continuos assessment from which the prediction is made is known as the **predictor** while the future task (performance in the examination) being predicted is called **criterion.**

The predictive validity of a predictor is determined by the statistical method. In the statistical approach, the predictor is correlated with the criterion using either Spearman Rank Order correlation or Pearson Product Moment formular. The correlation coefficient so obtained is called the predictive validity. Predictive validity is also called empirical validity.

Concurrent Validity: This type of validity shows the extent to which different tests of the same property are in agreement. It indicates how present performance could be used to estimate some other current measure of performance. When objectives test scores of a group are correlated with their essay scores, it is concurrent.

CONSTRUCT VALIDITY : A construct is a complex of mental images and impression systematically synthesized to aid the mind in further speculation. (Anikweze 2005). Construct validity indicates the extent to which two tests that measure conceptually related properties agree. When a test is used to describe the extent to which an individual possesses a psychological characteristics, the construct validity of the measure is the concern. The degree to which the presence of construct or characteristic in a testee is responsible for a score on a test mean construct validity. For example, tests of personality, mechanical aptitude, verbal ability, critical thinking and interest are validated in terms of their construct and the relation of their scores to pertinent external data.

TEST RELIABILITY AND TYPES

Test reliability is next to validity in order of importance. Reliability means the extent to which the test measures what ever it does measure in a consistent way. The test scores of students must be dependable and reproducible. It indicates the degree of accuracy with which a test measures what it designed to measure. It implies consistency of test results over time and item. It should be noted that a test might be reliable but not valid even though every valid test must also have the property of reliability.

There are three popular ways of estimating reliability. These are test – retest, parallel or equivalent form and split half.

1. **TEST – RETEST RELIABILITY**: Test retest reliability of a test is a measure if its stability over time. This type of reliability is determined by administering the same test twice to the same candidates under the approximately the same testing conditions given a time interval of between two to six weeks between the two administrations. The scores obtained from the two administrations are correlated using Person Product Moment correlation to obtain correlation coefficient called test retest reliability (or the coefficient of stability). The higher the coefficient the greater the stability.

PARALLEL OR EQUIVALENT FORM: This method of estimating reliability is based on the determination of consistency of performance by students across items that are intended to measure the same process objective. To solve problem of time interval, this method is developed. Two parallel or alternate forms of a test are administered concurrently to the same students. A correlation coefficient of the two sets of scores is computed to yield coefficient of equivalence

3. SPLIT – HALF RELIABILITY

The split – half method of determining test reliability is the most commonly employed because it is the easiest, involving only one test form, one group of testee, and one test administration (Obe 1980). The test is administered to the appropriate candidates and the resulting scores split into two equal halves usually odd and even numbers. The two half test scores are then correlated with one another. Pearson Product Moment

correlation method is used to obtain a split – half reliability coefficient called coefficient of internal consistency (rt). This rt describes the reliability of only half of the test having shortened the original test to half – length before correlating. To correct for this shortening. **SPEARMAN – BROWN'S PROPHECY**

Formular is applied thus

$$rtt = \frac{2rt}{1+rt}$$

where

rtt = reliability coefficient of internal consistency of the whole test

rt = coefficient of correlation between odd – half and even – half scores.

TEST USABILITY

A reliable and valid test could be unusable because of other factors that affect the quality of a good instrument apart from the validity and reliability of a test, another vital quality to look for is the usability. The usability of a test according to Obe (1980) refers to such practical consideration as:

i. Ease of administration.
ii. Time economy: to prepare, administer, and score.
iii. Financial economy: cost of testing materials etc.
iv. Norms: ease of interpretation and application.

Obe (1980) goes further by recommending the following precautions to ensure the usability of teacher – made tests:

1. The test should be typed, carefully proof – read, and duplicated to ensure a few copies in excess of the actual number of testee. To avoid leakage, duplication should be kept close to the day of testing as much as possible, all carbon and rough papers used and all rejected copies should be destroyed, questions should be deleted immediately after typing and printing and must not be saved in the computer and other security measure taken.
   Ease of administration may be ensured by making suitable sitting arrangement, giving clear test directions, using testing material which are safe and proper invigilation.
2. Time may be economized by planning the test so that the construction, administrations and scoring are not unduly time consuming.
3. The materials needed for the test should be readily available and not

too expensive.

4. Test norms (i.e. mean or typical students' scores, etc) are also desirable for important tests. Record of such norms will facilitate comparison of different individuals or set of students taking the same or similar tests from year to year.

CONCLUSION

This chapter has successfully defined the term test. The chapter also identified the various uses of test as listed by different authors. The chapter treated the common dimension of classifying test and its various characteristics. While it ended with properties of a good test.

REFERENCES

Adekunle, N.O. (2003). *Tests, Measurements and Educational Statistics.* Ibadan: The powerhouse press and publishers.

Anastasi, A. (1968). *Psychological Testing.* New York: Macmillan Limited.

Anikweze, C.M. (2005). *Measurement and Evaluation For Teacher Education.* Enugu: Snaap Press Ltd.

Cronbach, L.J. (1970). *Essentials of Psychological Testing.* New York: Harper and Row Publishers.

Drever, J. (1966). *A Dictionary of Psychology.* England: Penguin Books ltd.

Ibanga, J. (1981). *Guide on Tests & Measurement For Teachers and Students.* Calabar: Paico press and Books Ltd.

Gronhind, N.E. (1976). *Measurement and Evaluation in Teaching.* London: collier Macmillan Publishers.

Makinde, O. (1938). *Fundamentals of Guidance and counselling.* London: Macmillan Publishers.

Nwoye, N. (1990). *Counselling Psychology for Africa.* Jos: Fab Anieh (Nig) Ltd.

Obe, E.O. (1980). *Educational Testing in West Africa with Continuous Assessment.* Lagos: Premier press and publishers.

Thorndike, R.L. and Hegen, E. (1969). *Measurement and Evaluation in Psychology and Education.* New York: John Willey & Sons, Inc.